sciendo

Journal of Official Statistics, Vol. 36, No. 2, 2020, pp. 315–338, http://dx.doi.org/10.2478/JOS-2020-0017

Controlling for Selection Bias in Social Media Indicators through Official Statistics: a Proposal

Stefano M. Iacus[1], *Giuseppe Porro*[2], *Silvia Salini*[1], *and Elena Siletti*[1]

With the increase of social media usage, a huge new source of data has become available. Despite the enthusiasm linked to this revolution, one of the main outstanding criticisms in using these data is selection bias. Indeed, the reference population is unknown. Nevertheless, many studies show evidence that these data constitute a valuable source because they are more timely and possess higher space granularity. We propose to adjust statistics based on Twitter data by anchoring them to reliable official statistics through a weighted, space-time, small area estimation model. As a by-product, the proposed method also stabilizes the social media indicators, which is a welcome property required for official statistics. The method can be adapted anytime official statistics exists at the proper level of granularity and for which social media usage within the population is known. As an example, we adjust a subjective well-being indicator of "working conditions" in Italy, and combine it with relevant official statistics. The weights depend on broadband coverage and the Twitter rate at province level, while the analysis is performed at regional level. The resulting statistics are then compared with survey statistics on the "quality of job" at macro-economic regional level, showing evidence of similar paths.

Key words: Well-being; big data; sentiment analysis; small area estimation; weighting.

1. Introduction

Nowadays, researchers have potentially more data than ever before which has led to new progress in many fields of academia, government, industry, and commerce. However, although institutions and academics once had access to all the data produced because they collected or created it, have access to a smaller fraction of the data, since these data now locked up inside private companies. This information gap between the public and private sector requires further attention, but this discussion is outside of the scope of the present work. Social Networking Sites (SNS, or "social media") represent a special case for which a vast amount of data could be potentially accessible to public research.

Especially in the context of well-being measurement, the dramatic lack of timely data may be compensated by also considering alternative sources of data. SNS are a source of large and continuous flow of information, opinions, emotions, feelings and some researchers (Kwong et al. 2012; Hofacker et al. 2016) have considered them to be the largest available focus group in the world. The opinions expressed on SNS cover a large

[1] Department of Economics, Management and Quantitative Methods, University of Milan, Via Conservatorio 7 - 20122, Milan, Italy. Emails: stefano.iacus@unimi.it, silvia.salini@unimi.it, and elena.siletti@unimi.it
[2] Department of Law, Economics and Culture, Univertity of Insubria, Via Sant'Abbondio, 12 - 22100, Como, Italy. Email: giuseppe.porro@uninsubria.it.

spectrum of topics and interests, engage people from different social strata and usually do not suffer from censorship, although some exceptions have been investigated in King et al. (2013, 2014, 2017) .

Of course, these alternative sources of data are not, by design, intended to be used for the calculation of official statistics and in most cases, they are affected by different types of bias (Couper 2013). For example, in order to appear in the SNS data collections, individuals have to take some steps or satisfy some constraints like: have internet access (only 57% of the world population are in this set), open an account on the particular SNS targeted by the researchers for their analyses, and actively use it (about 45% of the world population are active users of SNS, see Table 1).

Other limitations come from the SNS themselves. No one can guarantee that these data will always exist in the future (we have seen the rise and fall of several platforms in recent years, changes of data structures and access policies). The use of public API (Application Programming Interface) or, even worse, web-scraping to obtain the data implies some lack of knowledge about the amount and quality of the data exposed by the SNS.

Despite the limitations that will be discussed in detail also in Section 3, there is a growing literature on social media as a source of data for preparing official statistics or composite indicators (see, e.g., Struijs et al. 2014; Culotta 2014; Daas et al. 2015; Alajajian et al. 2017; Tam and Clarke 2015; Kitchin 2015; Braaksma and Zeelenberg 2015; Severo et al. 2016; Van den Brakel et al. 2017) because nontraditional data are available at higher granularity, in time and space, compared to the data collected to produce official statistics.

In this article, we propose to extract emotions from social networks (Iacus et al. 2015, 2017) with the aim of building alternative subjective/perceived well-being indicators without directly surveying social network users, but only by interpreting their conversations on the internet. This approach of "listening" rather than "asking" has the potential advantage of getting rid of the nonresponse bias typical of surveys. The high-frequency rate of the data also allows taking into account that well-being is a mix of short-term, seasonal and long-term components.

Last, but not least, this article also proposes to address the selection bias problem of SNS indicators by anchoring them to official statistics and through the application of a space-time small area estimation (SAE) model (Rao 2005; Marhuenda et al. 2013) coupled with a weighting scheme.

The article is structured as follows: Section 2 introduces a multidimensional indicator of subjective well-being drawn from Twitter data: the Subjective Well-Being Index (SWBI). Section 3 discusses our proposal to control for sampling bias in Twitter-based indicators,

Table 1. Penetration data from the We Are Social and Hootsuite's report: "Digital in 2019" (Jan 2019), available at http://wearesocial.com; Annual digital growth from January 2018 to January 2019 in brackets.

Area	Internet users	Active social medial users
Global	57%(+9.1%)	45%(+9%)
European	86%(+7.6%)	55%(+3.2%)
Italian	92%(+27%)	59%(+2.9%)

combining a weighting scheme with a times-space SAE model. Section 4 restricts the focus to the component of the SWBI aimed at measuring the "quality of job/at work" and presents the results of an application of the method proposed in Section 3. Finally, Section 5 summarizes the conclusions of this work.

2. The SWBI: a Subjective Well-Being Index from Twitter Data

Since 2009, driven by the work of the Stiglitz Commission (Stiglitz et al. 2009), a large number of well-being indices have been developed – as alternatives or complements to traditional economic indicators, such as the GDP – with different structures, considering a great variety of dimensions, and for many purposes (Fleurbaey 2009). Generally, these new indicators come from survey data that, despite all efforts (Schwarz 1999; Schwarz and Strack 1999; Kahneman and Krueger 2006), still have some methodological drawbacks (Deaton 2011; Feddersen et al. 2016).

In particular, as Deaton (2011) pointed out, surveys are a potentially biased source of information; reports of well-being coming as answers to explicit questions may be influenced by contextual elements, such as the order of the questions or simply the fact that someone is asking for a personal well-being evaluation. The result is that information from surveys, as exemplified, is often subject to response error, in addition to the well-known nonresponse bias. Furthermore, surveys are costly and this makes it difficult to obtain data with a high time frequency or an adequate space granularity.

2.1. Sentiment Analysis and Twitter Data

As described in the introduction, SNS offer a large amount of data (Pentland 2014) that can be used for research purposes, enabling a new dimension of social dynamics study, as never before. Thanks to the progress of statistical methods for big data, social scientists are now able to manage and analyze data that are large in terms of dimensionality, size and time frequency (Lazer et al. 2009; King 2011). SNS like Twitter and Facebook, to mention a couple of them, have disclosed huge amounts of textual data and science shifted from traditional *text mining* to modern *sentiment/opinion analysis* with the aim of extracting semantic content from these types of data (Iacus 2014; King 2016).

The Integrated Sentiment Analysis (iSA) algorithm (Ceron et al. 2016) has been used in this work to construct a composite index of subjective well-being that attempts to capture various aspects of individual and collective life (Curini et al. 2015; Iacus et al. 2015). iSA is a human supervised machine learning method, in which a sample of texts (training set) is then first read and manually classified by human coders, and the rest of the corpus (test set) is automatically classified by the algorithm. The supervised part is essential, in that this is the step where qualitative information can be extracted from a text without relying on dictionaries or special semantic rules, but rather on cultural, psychological and emotional interpretation. Other approaches based on user-defined dictionaries exist, but mainly focus on the concept of happiness (Bollen et al. 2011; Zhao et al. 2018). The advantage of iSA over other machine learning techniques is that it is designed to directly estimate directly the aggregated distribution of the opinions (e.g., positive, negative, neutral) without passing through the individual classification of posts in the test set. This approach vastly reduces the estimation error. Moreover, as iSA is a

sequential method, in this context of highly noised data, the size of the training set needed to reach the same accuracy of other methods is usually smaller by a factor of 10 or 20 times. The reader can refer to Ceron et al. (2016) for the technical explanation of the method and its statistical properties.

It is important to note that the Twitter posts do not belong to individuals randomly chosen from a physical population (Baker et al. 2013; Murphy et al. 2014). The reference population is the population of posts of all Twitter accounts selected in the analysis. Moreover, Twitter accounts cannot be uniquely associated with individuals and some accounts are more active than others. For these reasons, the focus of our analysis is on the total volume of the posts collected (in Italy, during the reference period) through the public Twitter "search" and "streaming" API. These API are supposed to return a random sampling of the whole Twitter database, although by combining different strategies it is possible to get more. Comparing the volumes of the tweets we analyzed with the volumes obtained through a commercial provider, we could claim an almost similar coverage. However, for an institutional player, a commercial agreement should be considered as an alternative to our approach to data collection. A further restriction applies to our data set: only geo-referenced posts, about 1–5% of all tweets, have been collected. This further selection depends on individual Twitter users' privacy settings and hence may introduce additional bias. In our experience, if the analysis is based on geo-localized tweets at province level and the estimates are then aggregated at country level, the results are similar to those obtained on the whole set of tweets (with or without geo-reference information). From this personal and limited evidence, we can speculate, without any proof, that if this bias exists, it has a limited effect when data are aggregated at country level, but this is worth a further systematic investigation. To summarize, these data are clearly subject to selection bias arising in different ways: access to the internet,Twitter usage (not all people open and write on a Twitter account), Twitter platform API subsampling, and user specific privacy settings for geo-reference information. An attempt to deal with this overall bias will be presented in Section 3.

On the other hand, the advantage of using Twitter data is that the collection of data can be done in (almost) continuous time and in a wide range of sub-regional areas (in our case the Italian provinces). Finally, instead of asking something through a web form, thanks to the human supervised qualitative analysis, it is possible to capture expressions of well-being directly from the texts.

2.2. *The Construction of the SBWI Index*

The SWBI index (Iacus et al. 2015) is a multidimensional well-being indicator whose components were inspired by the dimensions adopted by the New Economic Foundation think-tank for its Happy Planet Index (New Economics Foundation 2012).

In summary, the SWBI consists of eight dimensions that concern three different well-being areas: personal well-being, social well-being, and well-being at work. More specifically,

1. *Personal well-being is defined as:*
    - **emotional well-being**: the overall balance between the frequency of experiencing positive and negative emotions, with higher scores showing that positive feelings are felt more often than negative ones (`emo`);

- **satisfying life**: having a positive assessment of one's life overall (`sat`);
- **vitality**: having energy, feeling well-rested and healthy while also being physically active (`vit`);
- **resilience and self-esteem**: a measure of individual psychological resources, of optimism and of the ability to deal with life stress (`res`); and
- **positive functioning**: feeling free to choose and having the opportunity to do it; being able to make use of personal skills while feeling absorbed and gratified in daily activities (`fun`).

2. *Social well-being is defined as:*
   - **trust and belonging**: trusting other people, feeling treated fairly and respectfully while experiencing sentiments of belonging (`tru`); and
   - **relationships**: the degree and quality of interactions in close relationships with family, friends and others who provide support (`rel`).
3. *Well-being at work is defined as:*
   - **quality of job**: feeling satished with a job, experiencing satisfaction with work-life balance, evaluating the emotional experiences of work and work conditions (`wor`).

The tweets written in the Italian language and posted from Italy constitute the SWBI's data source, and they were acquired via the public Twitter API. As mentioned, a share of the data (around 1% to 5%) includes geo-referenced information, which allows the estimation of the SWBI at a local level. As an experiment, in Iacus et al. (2019), the SWBI index has been estimated for the Italian provinces from 2012 to 2016 and compared to the "Il Sole 24 Ore" Quality of Life index (an indicator of life quality that is yearly evaluated and published by the "Il Sole 24 Ore" economic-financial newspaper in Italy).

Please note that, as SWBI does not use individual microdata, but is based on the aggregated sentiment analysis, it should be interpreted only as an aggregate measure of the level of well-being of a society.

3. A Proposal to Control for Bias in Social Media Estimates

In this section, we propose a method that makes use of official statistics to control the selection bias induced by the use of big social network data. In addition to a brief preamble to the basic SAE models, our approach, which is based on a weighted method and the SAE model, is discussed in what follows.

3.1. General SAE Models

SAE models play an important role in sampling theory and are employed when one needs to produce estimates in areas that are smaller than those for which the survey was planned. A *direct* estimator ($\hat{y}_d$), based only on the data coming from a limited-size sample from the small area, might be very unreliable; SAE *indirect* estimators are traditionally used to overcome this issue. Among indirect estimators, the model-based estimators are obtained by an explicit regression model, where a relationship between the target variable and some covariates is assumed. Model-based estimators can be classified as *unit-level* models, when

covariates are available at the unit level, and *area-level* models, when data are available only as area aggregates. In our case, as SWBI and official statistics exist only at province or regional level, the only option available is the area-level model.

The basic area-level model is the Fay-Herriot (FH) model (Fay and Herriot 1979), which is obtained as a linear mixed model in two stages consisting of a "sampling model" and a "linking model". Let $\hat{y}_d$ be a direct estimator of μ_d, a target unknown measure in area $d = 1, \ldots, D$: in the first stage, the "sampling model" (1) represents the uncertainty due to the fact that the target measure μ_d is unobservable and instead of it, only its measure on the sample $\hat{y}_d$ is available.

$$\hat{y}_d = \mu_d + e_d \tag{1}$$

$\hat{y}_d$ is unbiased, but unreliable, due to the small observed sample; and e_d are the sampling errors, which, given the characteristic of interest in d-th area, are assumed, for model convenience, to be independent and identically distributed (i.i.d.) with known variances, $N\left(0, \sigma_d^2\right)$.

In the second stage, the "linking model" (2) the area target measures μ_d are linearly related with a vector of area-level covariates $\boldsymbol{x}$.

$$\mu_d = \boldsymbol{x}_d'\boldsymbol{\beta} + u_d \tag{2}$$

where $\boldsymbol{\beta}$ is the common regression coefficients vector, and u_d are the model errors, un-observed and typically assumed i.i.d. from $N\left(0, \sigma_u^2\right)$. Combining the two model components (1) and (2), the hnal linear mixed model is dehned as follow:

$$\hat{y}_d = \boldsymbol{x}_d'\boldsymbol{\beta} + u_d + e_d \tag{3}$$

Several extensions of this basic area model have been proposed (Rao and Yu 1994; Ghosh et al. 1996; Singh et al. 2005; Marhuenda et al. 2013). Recently, these models have also been used with big data (Porter et al. 2014; Marchetti et al. 2015; Marchetti et al. 2016; Falorsi et al. 2017), which has been suggested for use as covariates when official statistics are either missing or poor. In particular, big data are used as covariates in area-level FH models, because these data are often unit level at the unit-level due to technical problems or legal restrictions. This is the case with social media search loads, remote sensing images or human mobility tracking.

Porter et al. (2014) used Google Trends searches as covariates in a spatial FH model, while in Falorsi et al. (2017), the time series query share from Google Trends was adopted as an auxiliary variable to improve the SAE model-based estimates for regional Italian youth unemployment. Marchetti et al. (2015) and Marchetti et al. (2016) have shown that big data improve the precision of small area estimates when used together with traditional covariates (i.e., official statistics or administrative data). More specifically, Marchetti et al. (2015) used big data as covariates in an FH model to estimate poverty indicators, accounting for the presence of measurement error, due to the availability of big data on mobility, using the Ybarra and Lohr (2008) approach. It is worth mentioning that Marchetti et al. (2015) suggested making use of survey data in some way to take into account the selection bias caused by the use of big data, but did not pursue this goal. This work is an attempt to implement their idea in a systematic way.

Marchetti et al. (2016) instead, used data coming from Twitter (Curini et al. 2015) as an instrumental covariate to estimate the Italian household share of food consumption expenditures at a provincial level, that is, they exploit the correlation between the official statistics indicator and social media data at regional level to reconstruct the official statistics at sub-regional level, thanks to the granularity of the Twitter data.

Conversely to the scholars cited above, in our proposal we do not use social media data (SWBI) as a covariate in a SAE model, but as a direct measure of the target unknown variable (well-being), and adopt official statistics as covariates in the area model. Following this goal, because social media data are biased, before applying the model we endorse a weighting procedure, as discussed in the next section.

3.2. *Weighting Strategy*

Usually, the methods adopted in the literature used to address the selection bias problem when using non-representative samples (e.g., the propensity score weighting (Rosembaum and Rubin 1983) or the Heckman correction (Heckman 1979)) are based on the use of unit level data (Cooper and Greenaway 2015). This also happens with social media data when individual characteristics of social media users are available. However, in light of the recently established privacy rules (GDPR) this is an increasingly remote eventuality. Note that, for Twitter data, the individual characteristics of every single account are not accurate or even unavailable and that SWBI is calculated as an aggregated estimate at province level. Unfortunately, as we will see later on, as the official statistics are available only at regional level, we adopt a hierarchical aggregation of the data at regional level, weighted by the characteristics of provincial macro-variables. As it will be explained via an application in Section 4, the macro-variables consist of the broadband coverage and the Twitter rate at provincial level. The aim is to take into account the selection bias that comes from the fact that not all people use or can use the internet and, among those who use the internet, not all of them make use of Twitter. The Twitter rate also compensates for the difference in Twitter volumes that we observe through the different geographical areas.

In particular, in Section 4, we consider $\hat{y}_{dt}^{w}$ as the regional sampling mean, where the regional units are the weighted means of province level units, in order to overcome the nonrandom sampling structure of the data:

$$\hat{y}_{dt}^{w} = \frac{1}{\sum_{i=1}^{n_{dt}} w_{idt}} \sum_{i=1}^{n_{dt}} y_{idt} w_{idt}, \tag{4}$$

where n_{dt} is the number of provinces in region d at time t, and w_{idt} are the weights. The choice of the actual weights depends on the application at hand. In Section 4, we will give a practical example. As an estimator of the variance of Equation (4), we adopt the plug-in estimator for weighted means:

$$\sigma^2_{\hat{y}^{w}_{dt}} = \frac{1}{n_{dt}} \left[\frac{1}{\sum_{i=1}^{n_{dt}} w_{idt}} \sum_{i=1}^{n_{dt}} y_{idt}^2 w_{idt} - \left(\hat{y}_{dt}^{w}\right)^2 \right]. \tag{5}$$

3.3. The Space-Time SAE Model with Weights

Since SWBI data are available for several periods of time T and domains D, we have chosen a particular SAE model, the spatio-temporal Fay-Herriot (STFH) model proposed by Marhuenda et al. (2013), to account for time and space correlations. This extension considers the spatial correlation between neighboring areas, while simultaneously including random effects for the time periods nested within areas. Thus, for domains $d = 1, 2, \ldots, D$ and time periods $t = 1, 2, \ldots, T$, let μ_{dt} be the target unknown measure (well-being) in area d at time t. The STFH model, just as any FH model, is composed of two stages. In the first stage, the "sampling model" is defined as:

$$\hat{y}^w_{dt} = \mu_{dt} + e_{dt}, \quad e_{dt} \overset{ind}{\sim} N\left(0, \sigma^2_{\hat{y}^w_{dt}}\right), \quad d = 1, 2, \ldots, D, \quad t = 1, 2, \ldots, T, \tag{6}$$

where e_{dt} are the sampling errors that are assumed to be independent and normally distributed, and $\sigma^2_{\hat{y}^w_{dt}}$ is an estimator of the variance as defined in Equation (5).

In the second stage of the STFH model, the "linking model" is as follows:

$$\mu_{dt} + \boldsymbol{x}'_{dt}\boldsymbol{\beta} + u_d + v_{dt} \quad u_d \overset{ind}{\sim} N\left(0, \sigma^2_1\right); \quad v_{dt} \overset{ind}{\sim} N\left(0, \sigma^2_2\right), \tag{7}$$

where $\boldsymbol{x}_{dt}$ is the column vector with the aggregated values of k covariates for the d-th area in t-th period of time and β is the vector of regression coefficients; u_d are the area effects that follow a first-order spatial autocorrelation process, SAR(1), with variance σ^2_1, spatial autocorrelation parameter ρ_1 and a $(d \times d)$ proximity matrix $\boldsymbol{W}$. Specifically, $\boldsymbol{W}$ is a row-standardized matrix obtained from an initial proximity matrix $\boldsymbol{W}^I$ whose diagonal elements are equal to zero and residual entries are equal to one, when the two domains are neighbours, and zero otherwise. Normality of u_d is required for the mean squared error, but not for point estimation. Furthermore, v_{dt} represents the area-time random effects that are assumed i.i.d. for each area d; these effects follow a first-order autoregressive process, AR(1), with the autocorrelation parameter ρ_2 and variance equal to σ^2_2. Accordingly, the final proposed linear mixed model is:

$$\hat{y}^w_{dt} = \boldsymbol{x}'_{dt}\boldsymbol{\beta} + u_d + v_{dt} + e_{dt}. \tag{8}$$

Therefore, $\boldsymbol{\theta} = \left(\rho_1, \sigma^2_1, \rho_2, \sigma^2_2\right)$is the vector of unknown parameters characterizing the spatio-temporal STFH model. Following Marhuenda et al. (2013), who provided $\hat{\boldsymbol{\beta}}$, the empirical best linear unbiased estimator (EBLUE) of $\boldsymbol{\beta}$, and $\hat{u}_d$ and $\hat{v}_{dt}$, the empirical best linear unbiased predictors (EBLUPs) of u_d and v_{dt}, are both obtained by replacing a consistent estimator $\hat{\boldsymbol{\theta}}$ in the respective BLUE and BLUPs introduced by Henderson (1975). The empirical estimation $\hat{\mu}_{dt}$ under the STFH model is given by:

$$\hat{\mu}_{dt} = \boldsymbol{x}'_{dt}\hat{\boldsymbol{\beta}} + \hat{u}_d + \hat{v}_{dt}. \tag{9}$$

As in Marhuenda et al. (2013), we use parametric bootstrap to estimate the mean squared error (MSE) of the EBLUPs. The MSE is calculated as follows:

$$MSE\left(\hat{\mu}_{dt}\right) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{u}^b_{dt} - \mu^b_{dt}\right)^2 \tag{10}$$

where, "b" remarks that these estimation is performed with the bootstrap procedure. And

$$\mu_{dt}^{b} = \boldsymbol{x}_{dt}'\hat{\boldsymbol{\beta}} + \hat{u}_{d}^{b} + \hat{v}_{dt}^{b}. \tag{11}$$

is the empirical estimation obtained in the first step of the bootstrap procedure using the bootstrap area and time effects: $\hat{u}_{d}^{b}$ and $\hat{v}_{dt}^{b}$.

In this way, the point estimate $\hat{\mu}_{dt}$ (indirect measure of well-being) of μ_{dt} (unknown well-being) can be supplemented with Equation (10) as a measure of uncertainty.

4. An Application to the Study of Well-Being at Work

The opportunity to integrate existing information on well-being with more information with a strong subjective and perceived trait, as those provided by social networks or specifically by SWBI, is a very interesting goal. In this section, with an application to Italian context we chose to use SWBI index and official statistics to guide our proposal. In particular, in Subsection 4.1, we describe the data that we use to implement the weighted procedure and the SAE model, and in Subsection 4.2 we discuss the result obtained.

4.1. Data and Variables

The SWBI index over 24 quarters from 2012 to 2017 is available at provincial and regional level. More than two hundred million tweets, in the period of the analysis were downloaded and classified, partly manually and partly through the iSA algorithm. The tweets have been classified as $+1$ (positive), 0 (neutral), or -1 (negative). The outcome variable is the estimated proportion of $+1$'s over the proportion of $+1$ and -1 and this represents the input variable y_{idt} in Equation (4).

As the variability of the number of tweets is remarkable, both along the time and the space dimension, there is the need to take into account this diversity. The range of data extends from a minimum of 1,727 tweets in 2016-Q1 for the Basilicata region to a maximum of 2,728,640 in 2017-Q2 for the Lombardia region. (Note that Valle d'Aosta has been dropped from the analysis as, considering that it consists of a single province, the proposed approach is not applicable because for example, random effects cannot be estimated.)

In order to have a more reliable view of the SWBI data at the regional level, we use the *Twitter rate* (i.e., the ratio between the number of tweets analysed and the population size in the area in the same period). The distribution of the Twitter rate over time among the Italian regions is shown in Figure 1. The average Twitter rate is around 18% ($SD = 12.29$), with a minimum regional value higher than 9% ($SD = 4.93$) in Campania, and a maximum regional value higher than 30% ($SD = 21.15$) in Molise (time averages for all the regions are blue points in the figure). The dispersion during the observational period is lower for large regions like Lazio, Puglia, Campania and Lombardia, and higher for small regions like Molise and Marche.

A better understanding of the SWBI information using the Twitter rate is made evident by examining Figure 2. The Twitter counts of 2017-Q4, shown on the left side of the figure, give the erroneous impression that most of the SWBI information comes from only a few large more populous regions (Piemonte, Lombardia, Veneto, Emilia and Campania),

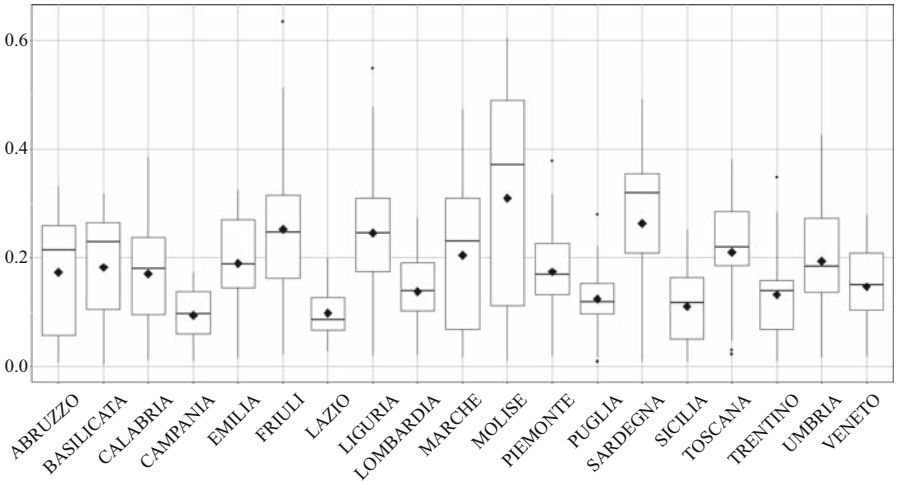

Fig. 1. Twitter rate for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.

while the Twitter rates displayed on the right side of the figure give the correct conclusion that all regions are homogeneously monitored.

4.1.1. The Construction of the Actual Weights

To implement the weighting procedure introduced in Subsection 3.2, after a selection process to define significant variables, we use the Twitter rate and the broadband coverage. The Twitter rate is closely related to mobile phone shares and broadband coverage is a measure of internet capacity. The use of these two variables is an attempt to take into account the selection bias. The Twitter rate, computed in each period and at province level,

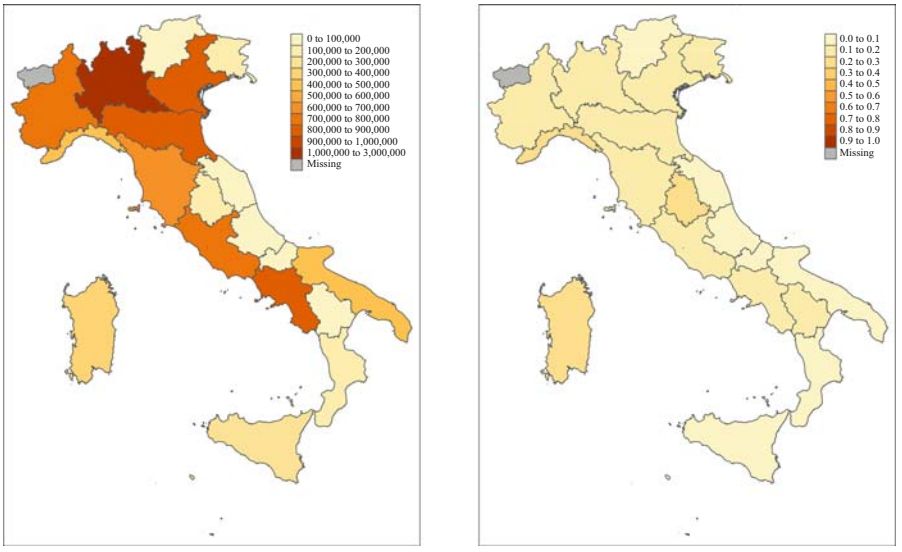

Fig. 2. Twitter counts map, on the left, and Twitter rates map, on the right, in the last quarter of 2014.

can be considered a good proxy of the use of Twitter for Italians. The broadband coverage is annual public data provided by *II Sole 24 Ore* and *Infratel Italia* for all the Italian provinces and can be considered the opportunity to access the internet in the different provinces. Coverage is quite stationary during a single year but, over time, what can happen is only an improvement of coverage in space or in signal intensity. Therefore, we replace the missing values with the data from the previous year to ensure that the coverage is not overestimated. The average broadband coverage is around 94% ($SD = 4.68$), with a minimum regional value of 72% ($SD = 4.57$) for Isernia in the Molise region. In 2012, the coverage mean was 92.15% ($SD = 3.9$) and in 2017, it was 92.65% ($SD = 5.6$). So, during the examined time period, the average broadband coverage remained the same, but the variability among regions increased, with an growth of around 42%. In detail, calling $w_{1;idt}$ the Twitter rate and $w_{2;idt}$ the broadband coverage, to apply to the weighting procedure for $\hat{y}^{w}_{dt}$ in Equation (4) and for $\sigma^{2}_{\hat{y}^{w}_{dt}}$ in Equation (5), we computed the weights as $w_{idt} = w_{1,idt} \cdot w_{2,idt}$.

4.1.2. Choosing the Covariates Among the Available Official Statistics

To apply the model proposed in Subsection 3.3, we need official statistics to use as covariates. After the Stiglitz's Commission suggestions, the Italian scenario of well-being measurement has increasingly changed. For example, the Italian National Institute of Statistics (ISTAT) set up the equitable and sustainable well-being project, where they plan a very complex system of well-being indicators, just following the same Commission suggestions. In 2013, they provided the BES ("Benessere Equo e Sostenibile", which, in English, is "Fair and Sustainable Well-being") index for the Italian regions, which analyses several dimensions of well-being.

Among these, the "work and life balance" dimension is the one that more closely relates to our research, although the construction of the composite indicator changed over time and it is not available for all quarters and provinces of Italy, making it impossible to use in our study.

ISTAT also provides other measures of well-being from the sample survey "Aspect of daily life"; however, these indicators are annual and representative for the five Italian macroeconomic areas: North-East, North-West, Center, South and Islands.

Discarding the idea to use the BES indexes and the "Aspect of daily life" survey measures, as covariates, we decided to rely on the only official statistics distributed by ISTAT that are available at least at the regional level and for the period of the analysis (although only for every quarter, the ISTAT data are available: http://dati.istat.it/ and http://demo.istat.it/). Despite the fact that the proposed model should work for each component of the SWBI at the province level, due to the limited availability of official statistics at frequencies higher than the year and at the sub-national level, we restrict our empirical analysis to the `wor` dimension of the SWBI. Even though the `wor` dimension could be monitored daily at province level, for the analysis they have been aggregated quarterly for each province ($\hat{y}_{idt}$).

The distribution of the unweighted `wor` ($\hat{y}_{dt}$) with regional aggregation over time is shown in Figure 3. The average of `wor` is 35.34% ($SD = 25.40$) with a minimum average regional value around 33% ($SD = 21.01$) in Sardegna and a maximum average regional value higher than 38% ($SD = 28.48$) in Lazio. The minimum and the maximum values of the quality of work are 9.01% for Lombardia in 2012-Q2 and 93.01% for Trentino in

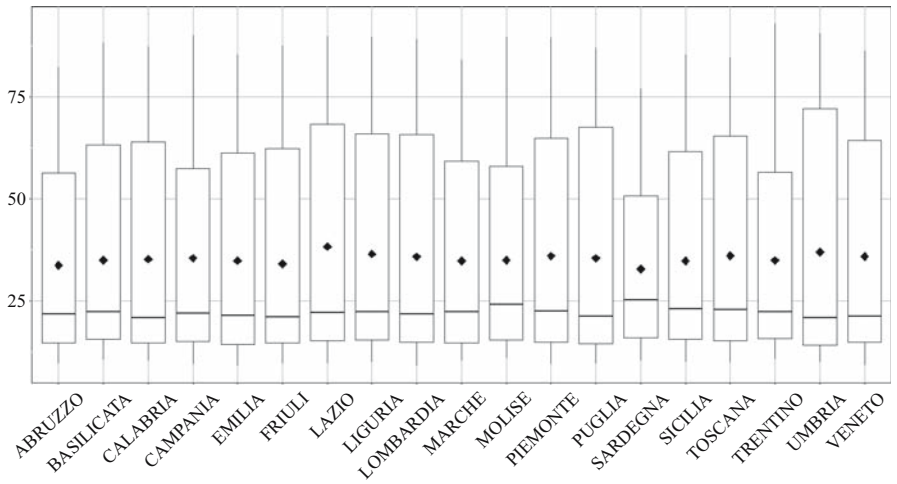

Fig. 3. The SWBI's unweighted wox *dimension ($\hat{y}_{dt}$) for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.*

2015-Q3, respectively. The similar averages are 35.79% ($SD = 26.87$) and 34.88% ($SD = 24.74$), respectively.

The considered area level auxiliary variables, before any process of selection, in the job context were as follows: the unemployment and inactivity rates, computed both in relation to the labour force (as they are traditionally calculated) and to the resident population; and the birth rate, the mortality rate and the natural rates, in the socio-demographic context. In the numerator of the natural rate there is the natural balance, which is the difference between births and deaths. After fitting the model, the selected covariates that make up the matrix $\boldsymbol{x}$ in Model (8), are the "unemployment rate" $\boldsymbol{x}_1$ and the "mortality rate" $\boldsymbol{x}_2$. The selection of these variables is the result of a standard model selection procedure after testing different variable configurations.

A large number of studies – since Clark and Oswald (1994) – provides documentary evidence of the negative relationship between unemployment and subjective well-being. It has also been argued that getting unemployed people back to work can do more for their well-being perception than subsidizing their unemployment status (see, e.g., Winkelmann 2014). In other words, non-pecuniary costs of unemployment are significant: therefore, higher unemployment rate (i.e., a higher risk of being unemployed) is here assumed to be related to the evaluation of well-being at work.

The relationship between working conditions and subjective well-being is often mediated, in the same literature, by health conditions: mortality or morbidity rates are assumed, in this respect, as proxies of health conditions.

The distribution of the unemployment rate over time among regions, as shown in Figure 4, reveals an average unemployment rate of 12.37% ($SD = 5.31$), with a minimum average regional value around 5% ($SD = 0.78$) for Trentino and a maximum average regional value higher than 22% ($SD = 2.13$) for Calabria. The same two regions also register the minimum and maximum values for the unemployment rate, 3.59% in 2017-Q3 and 25.15% in 2017-Q4, respectively.

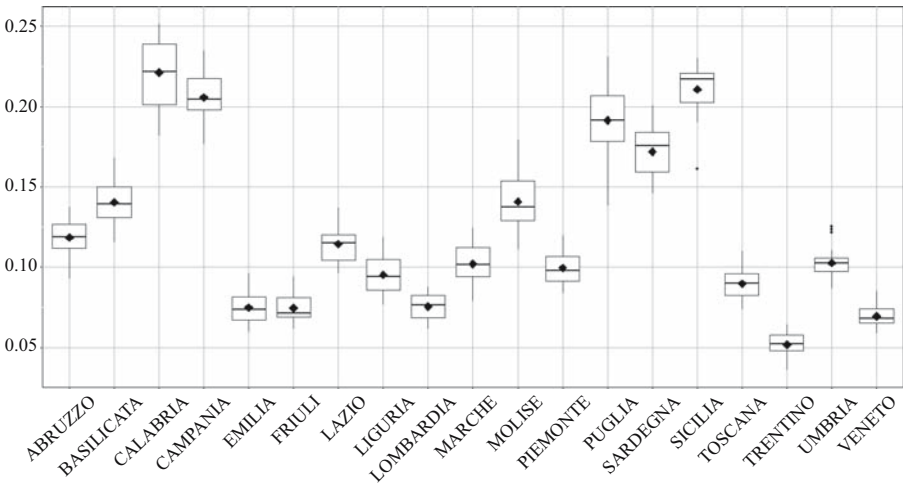

Fig. 4. Unemployment rate ($\mathbf{x}_1$) for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.

The distribution of the mortality rate over time among regions, as shown in Figure 5, illustrates an average mortality rate of 0.267% ($SD = 0.04$) with a minimum average regional value around 0.216% ($SD = 0.022$) in Trentino and a maximum average regional value higher than 0.343% ($SD = 0.032$) in Liguria. The same two regions also register the minimum and maximum values for the mortality rate, 0.19% in 2014-Q3 and 0.42% in 2017-Q1, respectively.

4.2. Results and Discussion

The weighted quality of job dimension $\hat{y}^w_{dt}$ (weighted `wor`), obtained following Equation (4), has remained stable with little variability between regions (Figure 6). The distributions

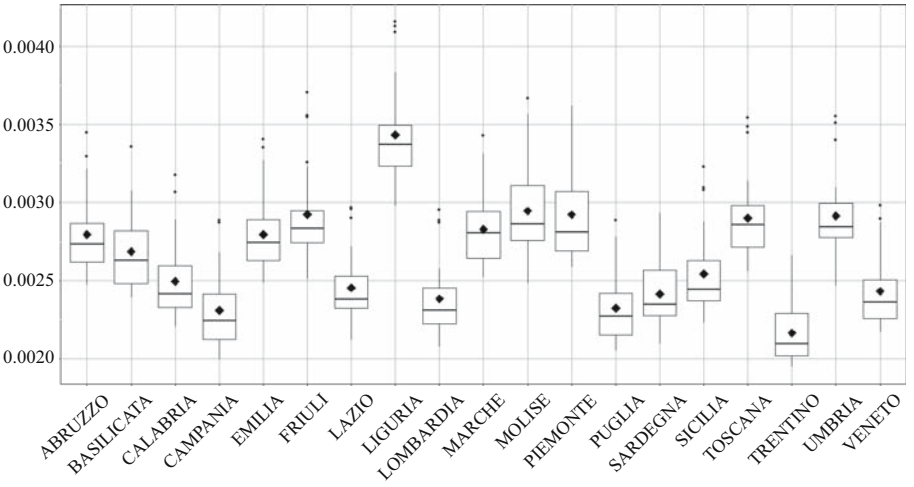

Fig. 5. Mortality rate ($\mathbf{x}_2$) for the considered Italian regions from the first quarter of 2012 to the last quarter of 2017.

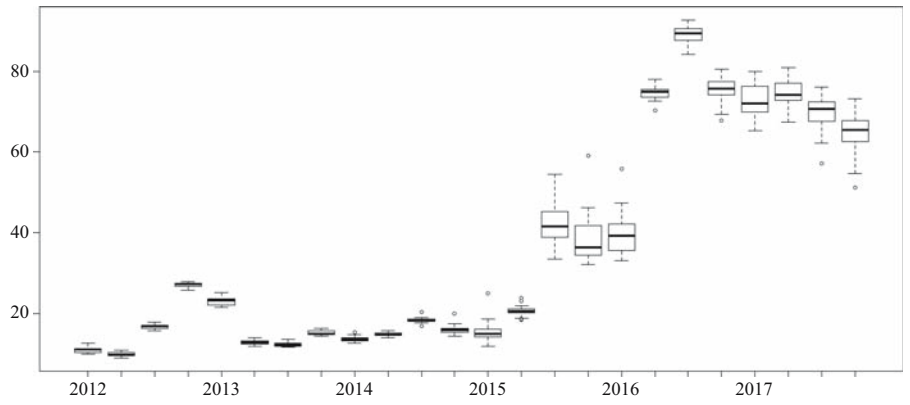

Fig. 6. The SWBI's weighted wor *dimension ($\hat{y}^w_{dt}$) during the periods from the first quarter of 2012 to the last quarter of 2017.*

were compressed until the second half of 2015, when they grew. This is especially evident from the second half of 2016, when this dimension attained values greater than 80, and even the differences between the regions were more marked, and the box-plots less crushed. Moreover, the average of $\hat{y}^w_{dt}$ is 36.17% ($SD = 26.38$) with a minimum average regional value around 34% ($SD = 22.91$) for Sardegna and a maximum average regional value higher than 39% ($SD = 29.24$) for Lazio, reflecting the earlier distributions shown in Figure 3 for $\hat{y}_{dt}$ (unweighted wor). The minimum and maximum values of the $\hat{y}^w_{dt}$ remained with Lombardia in 2012-Q2 (8.99%) and Trentino in 2015-Q3 (92.76%), respectively, and their averages were still similar (36.68% with $SD = 27.46$ for Lombardia and 37.99% with $SD = 28.66$ for Trentino).

Since comparing rankings is a valuable tool for policy makers and analysts, here we propose some discussions about them. The different rankings obtained by the two indices, both unweighted $\hat{y}_{dt}$ and weighted $\hat{y}^w_{dt}$, show no differences for around 4% of the cases ($\Delta =$ ranking differences), and only 15.6% of the cases show a Δ greater than four positions. The mean of the Δ is equal to 2.19 ($SD = 2.58$). Regions with the greatest differences were Trentino, Campania, Marche, and Sardegna, with the first two showing position improvement and the last two showing position weakening. For Trentino in particular, we remark that, after the weighting procedure, the greatest improvement took place during all four quarters of 2017.

In the applied STFH Model (8), data are available for $T = 24$ time instances, and the domains are $D = 19$, the considered Italian regions. Our data are "balanced" in that each region is measured using the same number of times and on the same occasions.

The row-standardized proximity matrix $\boldsymbol{W}_c$ of dimension 19×19 has been obtained from an initial proximity matrix, $\boldsymbol{W}^I_c$, whose diagonal elements are equal to zero and residual entries are equal to one, when the two regions had some common borders, and zero otherwise. Since in Italy, there are two regions corresponding to two islands (Sicilia and Sardegna), for these regions, we take other Italian regions with direct naval connections as neighbours.

As shown in Table 2, the coefficients for the covariates ($\hat{\beta}_1$ and $\hat{\beta}_2$) were both negative. This means that regions with larger unemployment and mortality rates had a poorer quality

Table 2. STFH model results.

(a) Estimated regression coefficients $\hat{\boldsymbol{\beta}}$ in Equation (9)

Variable	Coeff.	Std. Error	p-value
Intercept	62.72	5.49	0.000
Unemployment rate	−82.63	31.11	0.006
Mortality rate	−5649.48	1450.95	0.000

(b) Estimated vales for the vector of predictors $\hat{\boldsymbol{\theta}}$ and goodness of fit measures

Parameter	Estimate	Std. Error
$\hat{\sigma}_1^2$	0.0000	0.0000
$\hat{\rho}_1$	−0.0652	0.0000
$\hat{\sigma}_2^2$	94.72	0.0000
$\hat{\rho}_2$	0.8848	0.0000
Goodness of fit		
loglike	−1718.05	
AIC	3450.10	
BIC	3478.95	

of job dimension. The estimated spatial autocorrelation coefficient $\hat{\rho}_1$ is significant enough with a small negative value of about −0.07, (the size of the vector used is not large, $D = 19$), while the temporal autocorrelation coefficient $\hat{\rho}_2$ is still significant and has a greater positive value equal to about 0.88. The value equal to zero for $\hat{\sigma}_1^2$ is coherent with the analysis of distribution discussed above. The quality of job changes over time, but either little or not at all between regions.

4.2.1. The Weighted Measure of Well-Being at Work

In Figure 7, the scatter plots between the resulting $\hat{\mu}_{dt}$, obtained by fitting the STFH model, and the direct estimates, both unweighted $\hat{y}_{dt}$ (on the left) and weighted $\hat{y}_{dt}^w$ (on the right). In the SAE context, this graphical representation is used to test if the estimates are design unbiased: if the points lie along the diagonal, the direct estimates are approximately design unbiased, but if the points are under the line, the direct estimators are larger than the values predicted by the model, and vice versa if the points are above the line. Both the plots in the figure show points that lie along the diagonal for most of the cases. On the left side of the figure, we compare the SAE estimates $\hat{\mu}_{dt}$ with $\hat{y}_{dt}$, the unweighted estimates of `wor`, and there are more points away from the diagonal line than when the same estimated values are compared with $\hat{y}_{dt}^w$, the weighted estimates. Looking at the same plots, but for the different considered quarters, we find that the points away from the diagonal are in the periods where we have fewer analyzed tweets, and we observe an anomalous value of the variances. These two situations are caused by a lack of reliability in the information, but overall, we can conclude that the weighted estimates $\hat{y}_{dt}^w$ are approximately design unbiased.

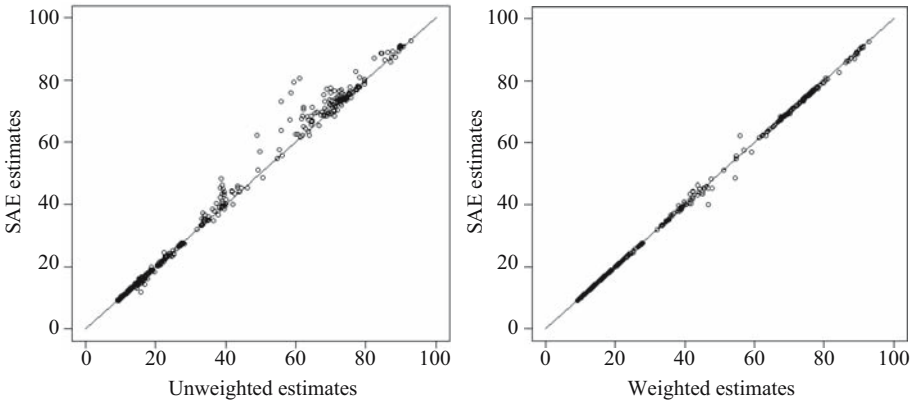

Fig. 7. Predicted values from the STFH model $\hat{\mu}_{dt}$ (SAE estimates) versus estimates of WOR, unweighted $\hat{y}_{dt}$ on the left and weighted $\hat{y}^{w}_{dt}$ on the right.

4.2.2. The Estimated Measure of Well-Being at Work from the Model

Considering the rankings, what changes if we use SAE model estimates instead of direct estimates, whether weighted or not?

Comparing the rankings obtained with the $\hat{y}_{dt}$ and those obtained with $\hat{\mu}_{dt}$,we find that in 29.2% of the cases the position is the same and in 15.8% of the cases, the Δ is greater than four. The mean of the ranking Δ is 2.16 ($SD = 2.58$). Equally, when we compared the above simple means $\hat{y}_{dt}$ with the weighted means $\hat{y}^{w}_{dt}$, regions with the greatest differences are Trentino, Campania, Marche, and Sardegna, with the first two showing position improvement and the last two showing position weakening. For Trentino, there is a great improvement during all quarters of 2017. Comparing the rankings obtained with the weighted values $\hat{y}_{dt}$ and those obtained with model estimates $\hat{\mu}_{dt}$ shows a very different situation: in 84.9% of the cases the positions are identical with less than 1% of the cases having a Δ greater than four (just one case has a great ranking difference: Marche in 2015-Q3 with a lag equal to eight positions). The average of the Δ equals 0.2 ($SD = 0.6$), which means that moving to weighted estimates $\hat{y}^{w}_{dt}$ with model predictions $\hat{\mu}_{dt}$ provides estimates that rank the same.

In SAE literature (Molina and Marhuenda 2015), coefficients of variations (CVs) are used traditionally to analyze the gain of efficiency for model estimates. While national statistical institutes are committed to publishing statistics with a high level of reliability, it is generally considered that estimates with CVs greater than 20% are not reliable. In Figures 9 and 10, the CVs of the three compared indices are shown, for the proposed final STFH model, and CVs were obtained by using the bootstrap procedure for the MSE estimates in Equation (10). As is evident, in our application, the CVs are always lower than 20%, except for fewer peaks. In particular, for most regions, the CVs are lower than 10% (Figure 10), while peak values are obtained in only a few quarters for 13 regions: Calabria, Campania, Emilia, Friuli, Lazio, Liguria, Marche, Molise, Piemonte, Lombardia, Sicilia, Toscana and Trentino. We stress that these high values of CVs are not stationary for these regions and it is clear that whenever we observe a peak of CVs, both the weighted indices and the model estimates improve reliability. Furthermore, CVs obtained for the model estimates ($\hat{\mu}_{dt}$, solid line) are always lower than the weighted estimations ($\hat{y}^{w}_{dt}$, dashed line)

Table 3. Pearson correlation coefficients r between ISTAT's WS and SAE-wor, in the five Italian geographical areas

Area	Overall	North-west	North-east	Central	South	Islands
r	0.245	0.694	0.383	0.581	0.849	0.480

and the unweighted estimates ($\hat{y}^w_{dt}$, dotted line). (For model estimates are computed as $CV = 100 \times \frac{\sqrt{MSE}}{Index}$, while for the others are $CV = 100 \times \frac{\sqrt{\text{Variance}}}{Index}$.)

Thus, values based on a STFH model look less variable in terms of the CV.

4.2.3. The Comparison Between the Estimated Measure of Well-Being at Work from the Model and an Official Index

In this section we compare our index obtained by the STFH model with an index of work satisfaction (WS) provided by ISTAT in its "Aspects of daily life" report. (All the details about the probability sample for the ISTAT survey "Aspects of daily life" can be found at www.istat.it/it/archivio/91926.)

The ISTAT's sample survey "Aspects of daily life" forms part of an integrated system of social surveys – The Multi-purpose Surveys on Household – and collects fundamental information on Italian individual and household daily life. It provides information on citizens' habits and the problems they face in everyday life. In the questionnaire, there are several thematic areas, based on different social aspects, that help describe the quality of individuals life, the degree of satisfaction of their conditions, their economic situation, the area in which they live, and the functioning of all public utility services, all topics

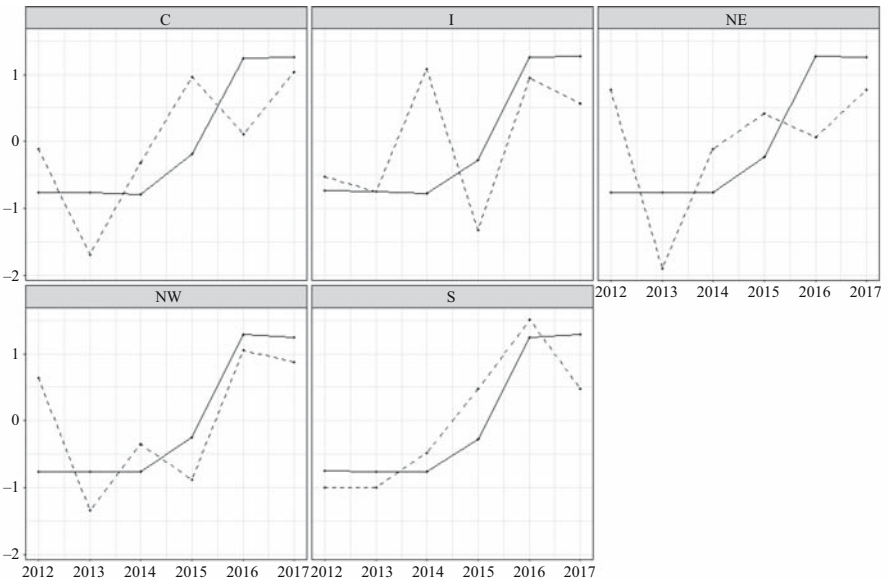

Fig. 8. Standardized time series of SAE-wor, solid line, and ISTAT's WS, dotted line, in the five Italian geographical areas (C: Central, I: Islands, NE: North-east, NW: North-west, S: South).

traditionally useful in studying the quality of life. This has been an annual survey since 2005, with data collection in February.

For our purpose we only consider WS, defined as the percentage of employed persons aged 15 years and over with a "good" level of satisfaction with their work. This index is computed as the sum of the percentages of people declaring to be "quite" and "very much" satisfied during the survey. Yearly WS data are distributed free of charge, but, as

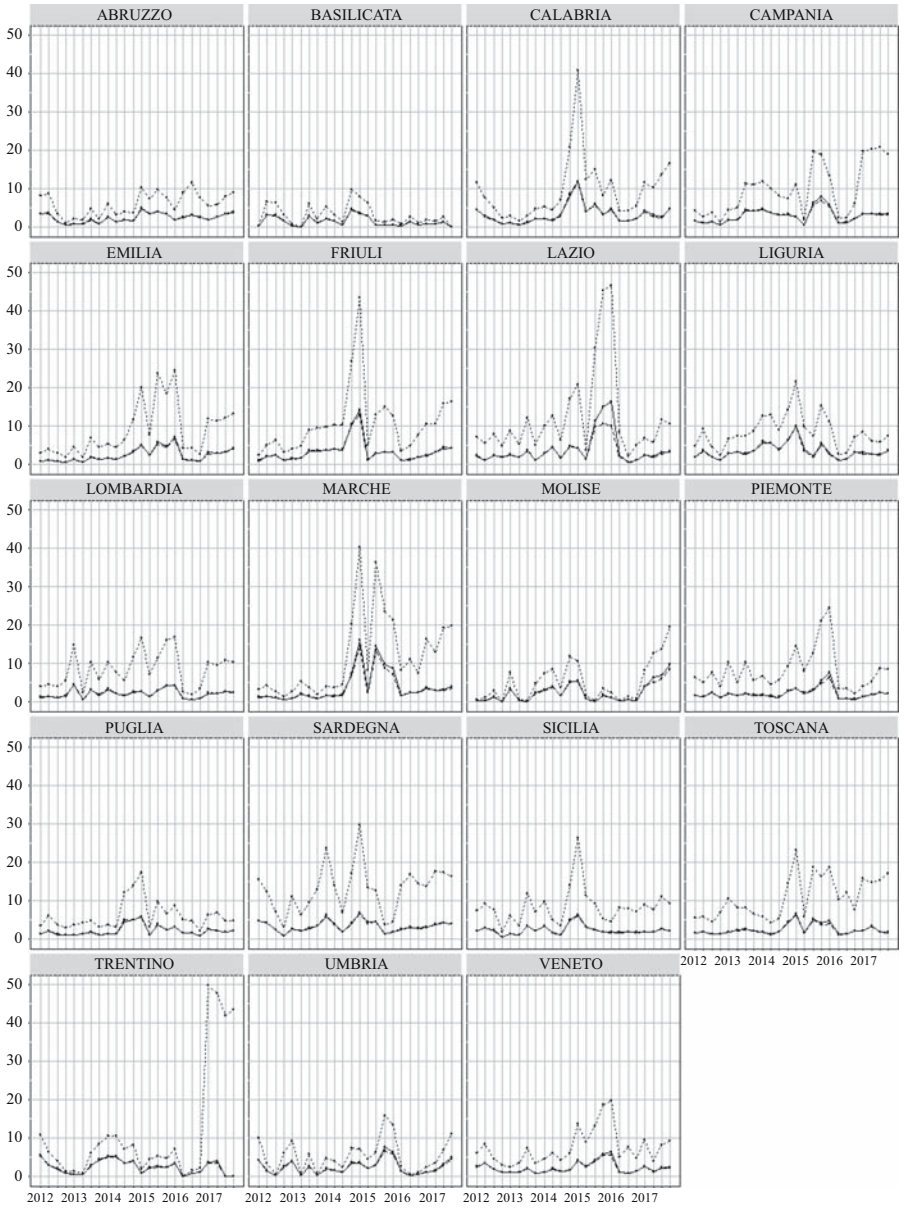

Fig. 9. Coefficient of variations for all the regions[2]; SAE estimates ($\hat{\mu}_{dt}$) with solid lines, weighted estimates ($\hat{y}^{w}_{dt}$) with dashed lines, and unweighted estimates ($\hat{y}_{dt}$) with dotted lines.

mentioned previously in the covariates section, they are representatives for the five Italian geographical areas: North-west, North-east, Central, South, and Islands.

To compare this index with our information, we aggregate the SAE estimates, $\hat{\mu}_{dt}$, obtained as discussed in the previous sections, yearly and in the same geographical areas, weighing with the corresponding resident population (SAE-`wor`).

The correlations between ISTAT index and SAE-`wor` are displayed in Table 3. If we consider all the overall data, the correlation is about 25%, while if we analyze the relationships within each area we find stronger links, with a maximum value in South Italy amounting to 85%.

Given the different scales of the ISTAT index and the proposed STFH estimator, for the purpose of visual comparison, Figure 8 represents the plot of their values, both standardized. Looking at these plots, the correlations become quite evident. We note that the correlation results are similar if we replace the STFH estimator with the raw `wor` measures (unweighted $\hat{y}^{w}_{dt}$ and weighted $\hat{y}^{w}_{dt}$).

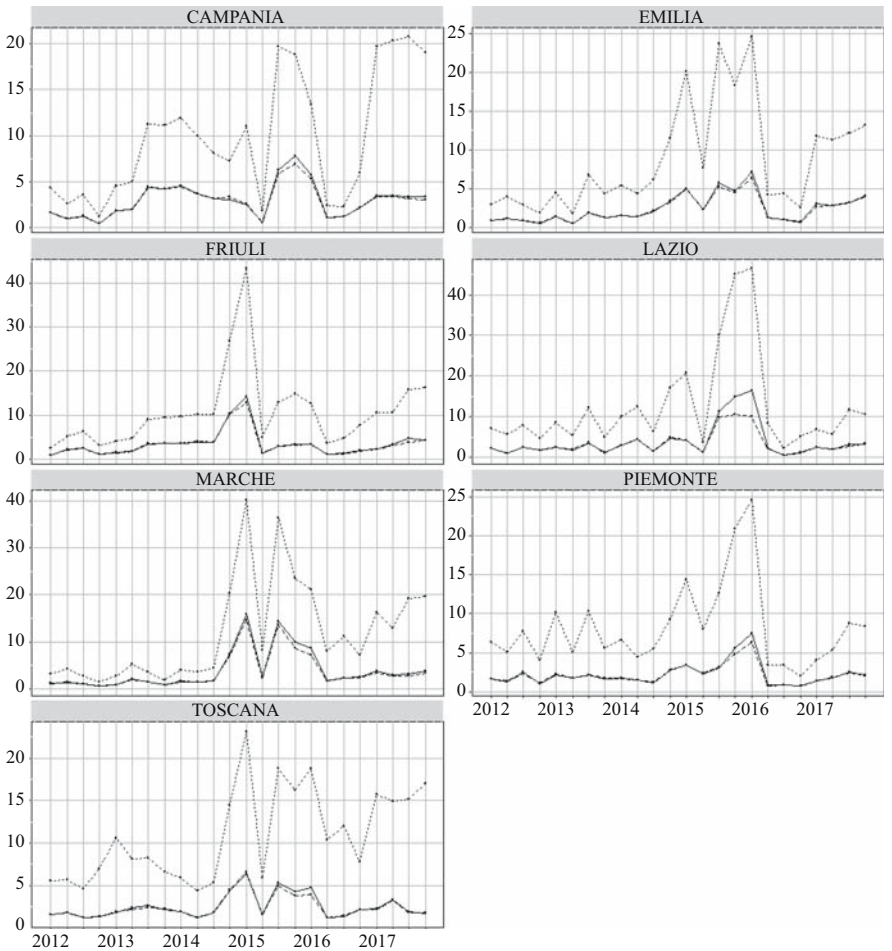

Fig. 10. Coefficient of variations for the regions with peaks greater than 20%[2]; SAE estimates ($\hat{\mu}_{dt}$) with solid lines, weighted estimates ($\hat{y}^{w}_{dt}$) with dashed lines, and unweighted estimates ($\hat{y}_{dt}$) with dotted lines.

5. Conclusion

The huge and increasing amount of data provided by social media is affected by selection bias that occurs either because not everyone has access to the internet or because not everyone who accesses the internet is interested in using social media. So far, this is a serious obstacle to using data from SNS for integrating into official statistics. To the best of our knowledge, there has been no systematic attempt to treat the bias problem, although we mentioned other important studies in which social media data have been considered along with official statistics and showed the added value in using this type of data.

In this article we have proposed to control selection bias caused by the use of aggregated data from social media by combining a weighting method and an SAE model.

Looking at the results, it seems that the selection bias inherent in social network data can be controlled using our approach. In particular, what we have shown is that – properly weighting statistics based on social media – we have approximately design unbiased statistics, that is, we have corrected the selection bias up to the only benchmark data available, which are the official statistics. We also gained additional properties through the SAE model, one of which is the stabilization of the variances of the social media statistics, which is a property required by official statistics. We have also shown that, despite using SNS data, the adjusted "wor" component of SWBI (albeit built upon different official statistics) correlates with the ISTAT statistics (available at macroeconomic level only) on the quality of work survey data.

This is clearly just the beginning of the story. Certainly, the accuracy of the proposed method could be improved using different SAE models based on dynamic systems so as to exploit fully the high resolution of the social media data, or by integrating more big data, sources at the same time, each with its own bias corrected statistics. These kinds of extensions represent interesting methodological challenges for the future.

6. References

Alajajian, S.E., J.R. Williams, A.J. Reagan, S.C. Alajajian, M.R. Frank, L. Mitchell, J. Lahne, C.M. Danforth, and P.S. Dodds. 2017. "The Lexicocalorimeter: Gauging public health through caloric input and output on social media." *PLOS ONE* 12(2)(February): 1–25. DOI: https://doi.org/10.1371/journal.pone.0168893.

Baker, R., J.M. Brick, N.A. Bates, M. Battaglia, M.P. Couper, J.A. Dever, K.J. Gile, and R. Tourangeau. 2013. "Summary Report of the AAPOR Task Force on Non-probability Sampling." *Journal of Survey Statistics and Methodology* 1(2): 90. DOI: https://doi.org/10.1093/jssam/smt008.

Bollen, J., B. Gonçalves, G. Ruan, and H. Mao. 2011. "Happiness is Assortative in Online Social Networks." *Artif. Life* (Cambridge, MA, USA) 17(3)(August): 237–251. DOI: https://doi.org/10.1162/artl_a_00034.

Braaksma, B. and K. Zeelenberg. 2015. "Re-make/Re-model: Should big data change the modelling paradigm in official statistics?" *Statistical Journal of the IAOS* 31(2): 193–202. DOI: https://doi.org/10.3233/sji-150892.

Ceron, A., L. Curini, and S.M. Iacus. 2016. "iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content." *Information Sciences* 367–368: 105–124. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2016.05.052.

Clark, A.E. and A.J. Oswald. 1994. "Unhappiness and Unemployment." *Economic Journal* 104(424): 648–659. DOI: https://doi.org/10.2307/2234639.

Cooper, D. and M. Greenaway. 2015. *Non-probability Survey Sampling in Official Statistics*. Office for National Statistics – Methodology Working Paper Series N4. Available at: https://www.k/ons/guide-method/method-quality/specific/gss-methodology-series/ons-working-paper-series/mwp3-non-probability-survey-sampling-in-official-statistics.pdf (accessed May 2020).

Couper, M.P. 2013. "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys." *Survey Research Methods* 7(3): 145–156. ISSN: 1864-3361. DOI: https://doi.org/10.18148/srm/2013.v7i3.5751.

Culotta, A. 2014. "Estimating County Health Statistics with Twitter." In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems,* 1335–1344. CHI '14. Toronto, Ontario, Canada: ACM. ISBN: 978-1-4503-2473-1. DOI: https://doi.org/10.1145/2556288.2557139.

Curini, L., S. Iacus, and L. Canova. 2015. "Measuring Idiosyncratic Happiness Through the Analysis of Twitter: An Application to the Italian Case." *Social Indicators Research* 121(2): 525–542. ISSN: 1573-0921. DOI: https://doi.org/10.1007/s11205-014-0646-2.

Daas, P.J.H., M.J. Puts, B. Buelens, and P. A.M. van den Hurk. "Big Data as a Source for Official Statistics." *Journal of Official Statistics* 31(2): 249–262. DOI: https://doi.org/10.1515/jos-2015-0016.

Deaton, A. 2011. "The Financial Crisis and the Well-Being of America." In *Investigations in the Economics of Aging,* edited by David A. Wise, 343–368. University of Chicago Press, June.

Falorsi, S., A. Fasulo, A. Naccarato, and M. Pratesi. 2017. *Small Area model for Italian regional monthly estimates of young unemployed using Google Trends Data*. 61st World Congress of the International Statistical Institute 16–21 July 2017 – Marrakech, Marocco, October. Available at: https://www.researchgate.net/publication/320554956_Small_Area_model_for_Italian_regional_monthly_estimates_of_young_unemployed_using_Google_Trends_Data (accessed May 2020).

Fay, R.E. and R.A. Herriot. 1979. "Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data." *Journal of the American Statistical Association* 74(366): 269–277. ISSN: 01621459. DOI: https://doi.org/10.2307/2286322.

Feddersen, J., R. Metcalfe, and M. Wooden. 2016. "Subjective wellbeing: why weather matters." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 179(1): 203–228. ISSN: 1467-985X. DOI: https://doi.org/10.1111/rssa.12118.

Fleurbaey, M. 2009. "Beyond GDP: The Quest for a Measure of Social Welfare." *Journal of Economic Literature* 47(4): 1029–1075. DOI: https://doi.org/10.1257/jel.47.4.1029.

Ghosh, M., N. Nangia, and D.H. Kim. 1996. "Estimation of Median Income of Four-Person Families: A Bayesian Time Series Approach." *Journal of the American Statistical Association* 91(436): 1423–1431. ISSN: 01621459. DOI: https://doi.org/10.2307/2291568.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1): 153–161. ISSN 00129682, 14680262. DOI: https://doi.org/10.2307/1912352.

Henderson, C.R. 1975. "Best Linear Unbiased Estimation and Prediction under a Selection Model." *Biometrics* 31(2): 423–447. ISSN 0006341X, 15410420. DOI: https://doi.org/10.2307/2529430.

Hofacker, C.F., E.C. Malthouse, and F. Sultan. 2016. "Big Data and consumer behavior: imminent opportunities." *Journal of Consumer Marketing* 33(2): 89–97. DOI: https://doi.org/10.1108/JCM-04-2015-1399.

Iacus, S.M. 2014. "Big Data or Big Fail?" The Good, the Bad and the Ugly and the missing role of Statistics. *Electronic Journal of Applied Statistical Analysis: Decision Support Systems and Services Evaluation* 5(1): 4–11. DOI: https://doi.org/10.1285/i2037-3627v5n1p4.

Iacus, S.M., G. Porro, S. Salini, and E. Siletti. 2015. "Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being." *ArXiv e-prints* Statistics – Applications (December): 1–26. Available at: 1512.01569 [stat.AP] (accessed December 2015).

Iacus, S.M., G. Porro, S. Salini, and E. Siletti. 2017. "How to exploit big data from social networks: a subjective well-being indicator via Twitter." In *SIS 2017. Statistics and data science: new challenges, new generations. Proceedings of the Conference of the Italian Statistical Society,* edited by Alessandra Petrucci and Rosanna Verde, 537–542. 28–30 June 2017, Firenze: Firenze University Press. ISBN: 978-88-6453-521-0

Iacus, S.M., G. Porro, S. Salini, and E. Siletti. 2019. "Social Networks Data and Subjective Well-Being. An Innovative Measurement for Italian Provinces." *Scienze Regionali, Italian Journal of Regional Science* Speciale (2019): 667–678. ISSN: 1720-3929. DOI: https://doi.org/10.14650/94673.

Kahneman, D. and A.B. Krueger. 2006. "Developments in the Measurement of Subjective Well-Being." *Journal of Economic Perspectives* 20(1): 3–24. DOI: https://doi.org/10.1257/089533006776526030.

King, G. 2011. "Ensuring the Data Rich Future of the Social Sciences." *Science* 331(February): 719–721. DOI: https://doi.org/10.1126/science.1197872.

King, G. 2016. "Preface: Big Data is Not About the Data!" Chap. 1 in *Computational Social Science: Discovery and Prediction,* edited by R. Michael Alvarez, 1–10. Cambridge: Cambridge University Press.

King, G., J. Pan, and M.E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2): 326–343. DOI: https://doi.org/10.1017/S0003055413000014.

King, G., J. Pan, and M.E. Roberts. 2014. "Reverse-engineering censorship in China: Randomized experimentation and participant observation." *Science* 345(6199): 891–913. ISSN: 0036-8075. DOI: https://doi.org/10.1126/science.1251722.

King, G., J. Pan, and M.E. Roberts. 2017. "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument." *American Political Science Review* 111(3): 484–501. DOI: https://doi.org/10.1017/S0003055417000144.

Kitchin, R. 2015. "The opportunities, challenges and risks of big data for official statistics." *Statistical Journal of the IAOS* 31(3): 471–481. DOI: https://doi.org/10.3233/SJI-150906.

Kwong, B.M., S.M. McPherson, J.F.A. Shibata, and O.T. Zee. 2012. "Facebook: Data mining the world's largest focus group." *Graziadia Business Review* 15: 1–8. Available at: https://gbr.pepperdine.edu/2012/11/facebook-data-mining-the-worlds-largest-focus-group/ (accessed April 2020).

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy, and M. van Alstyne. 2009. "Computational Social Science." *Science* 323(5915): 721–723. DOI: https://doi.org/10.1126/science.1167742.

Marchetti, S., C. Giusti, and M. Pratesi. 2016. "The use of Twitter data to improve small area estimates of households' share of food consumption expenditure in Italy." *AStA Wirtschafts – und Sozialstatistisches Archiv* 10(2)(October): 79–93. ISBN 1863-8163. DOI: https://doi.org/10.1007/s11943-016-0190-4.

Marchetti, S., C. Giusti, M. Pratesi, N. Salvati, F. Giannotti, D. Pedreschi, S. Rinzivillo, L. Pappalardo, and L. Gabrielli. 2015. "Small Area Model-Based Estimators Using Big Data Sources." *Journal of Official Statistics* 31(2): 263–281. DOI: https://doi.org/10.1515/jos-2015-0017.

Marhuenda, Y., I. Molina, and D. Morales. 2013. "Small area estimation with spatio-temporal Fay-Herriot models." The Third Special Issue on Statistical Signal Extraction and Filtering, *Computational Statistics & Data Analysis* 58: 308–325. ISSN: 0167-9473. DOI: https://doi.org/10.1016/j.csda.2012.09.002.

Molina, I. and Y. Marhuenda. 2015. "sae: An R package for small area estimation." *The R Journal* 7(1): 81–98. DOI: https://doi.org/10.32614/RJ-2015-007.

Murphy, J., M.W. Link, J. Childs, C. Tesfaye, E. Dean, M. Stern, J. Pasek, J. Cohen, M. Callegaro, and P. Harwood. 2014. "Social Media in Public Opinion Research Executive summary of the AAPOR task force on Emerging Technologies in Public Opinion Research." *Public Opinion Quarterly* 78(4): 788–794. DOI: https://doi.org/10.1093/poq/nfu053.

New Economics Foundation. 2012. *The Happy Planet Index: 2012 Report. A global index of sustainable well-being*. New Economics Foundation. Available at: https://neweconomics.org/uploads/files/d8879619b64bae461f_opm6ixqee.pdf (accessed August 2015).

Pentland, A. 2014. *Social Physics: how good ideas spread – the lessons from a new science*. EBL-Schweitzer. Scribe Publications Pty Limited. ISBN: 978113143.

Porter, A.T., S.H. Holan, C.K. Wikle, and N. Cressie. 2014. "Spatial Fay-Herriot models for small area estimation with functional covariates." *Spatial Statistics* 10: 27–42. DOI: https://doi.org/10.1016/j.spasta.2014.07.001.

Rao, J.N.K. and M. Yu. 1994. "Small-Area Estimation by Combining Time-Series and Cross-Sectional Data." *The Canadian Journal of Statistics* 22(4): 511–528. ISSN: 03195724. DOI: https://doi.org/10.2307/3315407.

Rao, J.N.K. 2005. *Small Area Estimation.* Wiley Series in Survey Methodology. John Wiley & Sons, January. ISBN: 9780471431626.

Rosembaum, P.R. and D.B. Rubin. 1983. "The central role of the propensity score in observational studies for causal effects." *Biometrika* 70(1): 41–55. DOI: https://doi.org/10.2307/2335942.

Schwarz, N. 1999. "Self-reports: how the questions shape the answers." *American psychologist* 54(2): 93–105. DOI: https://doi.org/10.1037/0003-066X.54.2.93.

Schwarz, N. and F. Strack. 1999. "Reports of subjective well-being: Judgmental processes and their methodological implications." In *Well-being: The foundations of hedonic psychology,* edited by D. Kahneman, E. Diener, and N. Schwarz, 7: 61–84. New York: Russell Sage Foundation.

Severo, M., A. Feredj, and A. Romele. 2016. "Soft Data and Public Policy: Can Social Media Offer Alternatives to Official Statistics in Urban Policymaking?" *Policy & Internet* 8(3)(September): 354–372. ISSN: 1944-2866. DOI: https://doi.org/10.1002/poi3.127.

Singh, B.B., G.K. Shukla, and D. Kundu. 2005. "Spatio-temporal models in small area estimation." *Survey Methodology* 31(2): 183–195. DOI: https://doi.org/10.1.1.617.1513.

Stiglitz, J., A. Sen, and J.-P. Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*. INSEE. Available at: https://www.researchgate.net/publication/258260767_Report_of_the_Commission_on_the_Measurement_of_Economic_Performance_and_Social_Progress_CMEPSP (accessed April 2020).

Struijs, P., B. Braaksma, and P.J.H. Daas. 2014. "Official statistics and Big Data." *Big Data & Society* 1(1): 1–6. DOI: https://doi.org/10.1177/2053951714538417.

Tam, S.-M. and F. Clarke. 2015. "Big Data, Official Statistics and Some Initiatives by the Australian Bureau of Statistics." *International Statistical Review* 83(3)(December): 436–448. DOI: https://doi.org/10.1111/insr.12105.

Van den Brakel, J., J. Söhler, P.J.H. Daas, and B. Buelens. 2017. "Social media as a data source for official statistics; the Dutch Consumer Conhdence Index." *Survey Methodology* 12-001-X (43): 183–210. DOI: https://doi.org/10.13140/RG.2.2.19294.64326.

Winkelmann, R. 2014. "Unhappiness and Unemployment." *IZA World of Labor* 94. DOI: https://doi.org/10.15185/izawol.94.

Ybarra, L.M.R. and S.L. Lohr. 2008. "Small Area Estimation When Auxiliary Information Is Measured with Error." *Biometrika* 95(4): 919–931. ISSN: 00063444. DOI: https://doi.org/10.1093/biomet/asn048.

Zhao, Y., F. Yu, B. Jing, X. Hu, A. Luo, and K. Peng. 2018. "An Analysis of Well-Being Determinants at the City Level in China Using Big Data." *Social Indicators Research* (October). ISSN: 1573-0921. DOI: https://doi.org/10.1007/s11205-018-2015-z.

Received March 2019
Revised July 2019
Accepted January 2020